Universidade Estadual de Campinas
Instituto de Computação

Plano de Desenvolvimento da Disciplina

MC959 - Tópicos em Inteligência Artificial
Oferecimento: Aprendizado de Máquina Ético
Docente: Prof. Marcos M. Raimundo

Descrição

Este documento descreve, de forma sucinta, o plano de desenvolvimento da disciplina (PDD) de MC959 - Tópicos em Inteligência Artificial. Em particular, são destacados, de acordo com os requisitos do Regimento Geral de Graduação, o cronograma de atividades, os critérios de avaliação, punição para fraudes e plágios e a bibliografia a ser utilizada ao longo do semestre.

1. Programa da disciplina

O programa desta turma da disciplina cobrirá, em diferentes níveis de profundidade, tópicos relacionados à ética em aprendizado de máquina. O objetivo terminal da disciplina é que o aluno seja ser capaz de: a) identificar o uso de dados e de ferramentas de aprendizado de máquina que gerem danos a indivíduos e grupos vulneráveis e, b) utilizar e propor ferramentas que reduzam o impacto e aumentem a transparência de tais dados e ferramentas. Para isso, serão abordados os seguintes assuntos:

- Ética, Confiança e Segurança: Conceitos básicos para produzir melhores ferramentas inteligentes.
- Gerenciamento de dados: Fontes de dados, modalidades, vieses, privacidade e consentimento, Leis de proteção.
- Modelagem básica de modelos de aprendizado: Conceitos básicos de aprendizado de máquina, aprendizado supervisionado, modelos causais
- Confiabilidade e Imparcialidade: Detecção e promoção de imparcialidade em máquinas de aprendizado, manutenção de confiabilidade em cenários sob mudança ou ataque.
- Interação: Interpretabilidade em aprendizado de máquina, e criação de mecanismos de intervenção nas máquinas de aprendizado.
- Propósito: Bem social, influência dos tomadores de decisão no desenho de produtos de aprendizado de máquina, desafios em aprendizado de máquina.

Cada assunto será coberto com um determinado número de aulas, que deverá variar entre duas e seis aulas.

2. Cronograma de atividades

Esta disciplina não terá exames individuais, e os alunos serão avaliados através de 6 (seis) trabalhos realizados em trios. Os trabalhos em trio avaliarão o domínio dos alunos dos conceitos e técnicas aprendidos em aula, através de exercícios práticos e redação de relatórios técnicos (a entrega dos relatórios e códigos-fonte usando Jupyter são mandatórias). A aplicação norte desses trabalhos será escolhida pelos alunos, e as soluções aplicadas a esses problemas/bases de dados deverão ser implementadas em Python, utilizando Jupyter Notebook. As soluções poderão utilizar bibliotecas de Aprendizado de Máquina tal como a scikit-learn, quando não puderem ser usadas, tal informação será explicita no enunciado. Os enunciados de cada trabalho e suas respectivas datas de entrega serão distribuídos de forma aproximadamente uniforme ao longo do semestre letivo. Além disso, a disciplina contará com um trabalho para avaliação diagnóstica e um trabalho final ao lugar do exame. Ambas avaliações não são mandatórias, e só poderão aumentar a nota final.

3. Critérios de avaliação

Cada um dos trabalhos $i \in \{1, 2, 3, 4, 5, 6\}$ com nota N_i (a ser atribuída) terá ponderação P_i e será avaliado através de conceitos, de A a F. As ponderações serão $\{2, 3, 3, 7, 7, 2\}$ para os trabalhos de 1 a 6 respectivamente. A nota final NF consiste na "mediana ponderada" das notas N_i. Em outras palavras, NF consiste na mediana do conjunto de avaliação $\mathcal{M}$. O número de elementos m_c de cada conceito $c \in \{A, B, C, D, Z\}$ no conjunto $\mathcal{M}$ corresponde a soma das ponderações para cada trabalho que tenha sido atribuído aquele conceito [1].

$$m_c = \sum_{i:T_i=c} P_i \tag{1}$$

O sistema de conceitos irá se basear na qualidade da execução de cada projeto proposto. Seguindo os seguintes nortes:

A - Execução exemplar (raros equívocos), com pesquisa adicional (ou inovações).

B - Execução exemplar (raros equívocos), sem pesquisa adicional (ou inovações). Execução com falhas, mas com pesquisa adicional (ou inovações).

C - Execução com falhas e sem pesquisa adicional (ou inovações).

D - Execução com falhas excessivas ou completamente incoerente.

Z - Não execução.

A avaliação diagnóstica terá peso 1, e o trabalho final terá peso 25, estas avaliações só serão contabilizadas caso melhorem o conceito final. O mapeamento entre conceito final e as notas numéricas será feito da seguinte forma: A - 10, B - 8, C - 6, D - 3, Z - 0.

Cláusulas:

- Cláusula 1: Um aluno com o conceito A será rebaixado para B se tiver mais que 1/4 dos trabalhos com conceito abaixo de E. E um aluno com conceito B será rebaixado para C se tiver mais que 1/2 dos trabalhos com conceito D.

[1] Exemplo: suponto as ponderações $P = \{1, 2, 3, 2\}$, e um aluno tirou $N = \{B, A, C, B\}$. O conjunto de avaliação é dado por $M \equiv \{C, C, C, B, B, B, A, A\}$, logo sua mediana é dada por B.

- Cláusula 2: Em caso de discrepância de esforço nas entregas de trabalho o(s) estudantes poderão, em acordo, solicitar readequação de notas. Pode ser solicitado o aumento de um conceito para cada decréscimo de um conceito dentro do grupo. (Ex: se um grupo de 3 anulos tirou B, um aluno pode ter sua nota reduzida (para C) para que outro tenha a nota aumentada (para A))

4. Punição para fraudes e plágios

Os trabalhos em trio devem ser resolvidos apenas pelos componentes da mesma, sem a consulta de terceiros. Detecção de fraude ou plágio em um trabalho implica em $N_i = 0$ para todos os envolvidos (quem recebeu ajuda e também quem ajudou). Reincidência implica em $NF = 0$ para todos os envolvidos.

5. Bibliografia a ser utilizada

O docente se baseará fortemente na referência "Trustworthy Machine Learning" para cumprir o cronograma proposto na Seção 1; abaixo segue essa e outras referências-base para a disciplina. Além disso material didático sobre os assuntos listados na Seção 1 poderão ser retirados de artigos e e outros livros.

- "Trustworthy Machine Learning" [4].
- "Fairness and machine learning" [1].
- "Practical Fairness" [3].
- "Interpretable Machine Learning" [2].

Referencias

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2, 2017.

[2] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[3] Aileen Nielsen. *Practical Fairness*. O'Reilly Media, 2020.

[4] Kush R Varshney. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3):26–29, 2019.